IN THE CLAIMS:

The following claim listing replaces all previous claim listings. Please amend claims 1 as follows:

Claim 1. (currently amended) A method for executing a network-based distributed application, the method comprising:

executing application instances of the distributed application ~~in~~ by application containers, each application container sharing state information about its application instance with other application containers;

calculating quality of service metrics for each application instance by the application containers; and

distributing application workload among the application instances using a decentralized workload management layer based on the quality of service metrics.

Claim 2. (original) The method of claim 1, further comprising associating application containers with autonomous workload management elements, the workload management elements forming the workload management layer.

Claim 3. (original) The method of claim 2, further comprising coordinating the application instances through a coordination mechanism coupled to the workload management layer.

Claim 4. (original) The method of claim 1, wherein distributing application workload among the application instances further comprises reducing workload assigned to an application container when the quality of service metrics reach an overload threshold value.

Claim 5. (currently amended) The method of claim 4, wherein reducing workload assigned to the application container further comprises:

examining an encoding of work unit groups provided by each application instance;

splitting a currently assigned work unit group into smaller work unit

groups;

assigning at least one of the smaller work unit groups to other application containers; and

utilizing a coordination mechanism to update changes in workload assignments to the other application containers.

Claim 6. (original) The method of claim 1, wherein distributing application workload among the application instances further comprises increasing workload assigned to the application container when the quality of service metrics reach an under-load threshold value.

Claim 7. (original) The method of claim 6, wherein increasing workload assigned to the application container further comprises:

examining an encoding of work unit groups provided by each application instance;

combining at least two currently assigned work unit groups into a smaller work unit group;

assigning the smaller work unit group to the application container; and

utilizing a coordination mechanism to update changes in workload assignments to the other application containers.

Claim 8. (original) The method of claim 1, further comprising dividing workload assigned to a single application instance to at least two application instances if a quality of service metric reaches an overload threshold.

Claim 9. (original) The method of claim 1, further comprising:

dividing a total workload performed by the distributed application among the application instances;

assigning each of the application instances a fractional workload; and

filtering client requests at the application containers based on the fractional workload assigned to the application instances.

Claim 10. (original) The method of claim 9, further comprising

migrating a client from a first application container to a second application container if workload from the client is not assigned to the application instance executing at the first application container.

Claim 11. The method of claim 10, further comprising labeling client requests such that application containers can determine if the requests belong to the fractional workload assigned to the application instances.

Claim 12. (original) The method of claim 1, further comprising receiving the application instances from application loaders.

Claims 13-32 (canceled)